# Combinatory Usability Evaluation of an Educational Virtual Museum Interface

Karoulis A.[1], Sylaiou S.[1], White M.[2]
[1] Aristotle University of Thessaloniki, Greece
[2] University of Sussex, U.K.
karoulis@csd.auth.gr, sylaiou@photo.topo.auth.gr, M.White@sussex.ac.uk

**Abstract**

*This paper presents the usability evaluation study that has been undertaken for the Augmented Representation of Cultural Objects (ARCO) system. The main purpose of this system is to integrate an enhanced educative and entertaining experience to virtual museum visitors. The aim of the current research is the evaluation of the interface of the system. Users as well as domain experts were recruited to investigate the most effective combination of user-based and expert-based evaluation, in order to elicit the most valuable results. Quantitative as well as qualitative approaches have also been employed, thus providing a framework for a holistic evaluation of the usability of an interface of such kind.*

## Introduction

Current research and an extensive survey to European museum sector ([4]; [14]) have shown that the World Wide Web enhanced by 3D visualization tools, such as the promising Virtual Reality (VR), Augmented Reality (AR) and Web3D technologies in conjunction with database technology, may facilitate the preservation, dissemination and presentation of cultural artefacts in museums' collections and also educate in an innovative and attractive way the wide public.

This study concerns the ARCO (Augmented Representation of Cultural Objects) [1] system, which integrates both commercial components and international standards and harnesses the potential of the World Wide Web. It allows museums curators to create digital artefacts, manage and build virtual museum exhibitions and publish them to the World Wide Web or to museum informative kiosks. The visualization of cultural objects consists of Web pages with virtual museum exhibitions that have embedded 3D VRML (Virtual Reality Markup Language) objects or/and 3D galleries, where objects can be browsed while walking in a 3D room, which is a reconstruction of a real gallery.

However, ARCO system had to be evaluated, not only with a demonstration of its capacities, but also through the contribution of real end-users. Two different groups of users, the domain specialists (museum curators) and simple end-users participated and evaluated various aspects of one component of the ARCO system, namely the ARIF (Augmented Reality Interface) component, which is the interface the end user will come in touch with. By means of interviews and structured questionnaires, appropriate information was collected, in order not only to improve the system characteristics, but also to understand if the system is usable, enjoyable, meaningful and appealing to the users.

## The ARCO System

The ARCO system allows museum curators to build, manage, archive and present virtual museum exhibitions based on 3D models of artifacts. ARCO also allows end-users to explore virtual exhibitions implemented using the system, and is introduced and described in detail in [16]. The 3D models are accompanied by images, texts, metadata information, sounds and movies. The cultural artifacts are digitized by means of a custom built stereo photogrammetry system (Object Modeler), mainly for digitizing small and medium sized objects and a custom modeling framework (Interactive Model Refinement and Rendering tool) that is used, in order to refine the digitized artifact [12]. These virtual reconstructions are managed through the use of a specially designed ARCO Content Management Application, which also allows the museum to build and publish virtual exhibitions to the Internet or a museum kiosk system.

Two main components of the ARCO system were of interest for evaluation: the ARCO Content Management Application (ACMA) and Augmented Reality Interface (ARIF). ACMA allows publishing of virtual museums to both Web and a specially designed application (ARIF) for switching between the Web and an AR system. The ARIF component is a presentation or visualisation framework that consists of three main subcomponents:

- The *ARIF Exhibition Server*. Data stored in the ARCO Database is visualized on user interfaces via the ARIF Exhibition Server.
- The *ARIF Presentation Domains* with implemented web browser functionality, suited for web-based presentations.
- The *ARIF AR—Augmented reality functionality*. This sub-component provides an AR based virtual

museum exhibition experience on a touch screen in the museum environment using table-top AR learning experiences, e.g. AR quizzes for educative purposes and on-line museum exhibitions.

## Usability Evaluation

According to [5] (Ergonomic requirements for office work with visual display terminals) ISO-standard, we have the following definition of usability: *Usability of a system is its ability to function effectively and efficiently, while providing subjective satisfaction to its users.*

Two important conceptions regarding the usability of an interface are "transparency" and "intuitiveness" [10]; [13]. Transparency refers to the ability of the interface to fade out in the background, allowing the user to concentrate during his work on *what* he wants to do and not on *how* to do it, in our case not interfering with the learning procedure, while intuitiveness refers to its ability to guide the user through it by the use of proper metaphors and successful mapping to the real world, e.g. by providing him with the appropriate icons, correct labeling, exact phrasing, constructive feedback etc.

The most applied methodologies are the expert-based and the empirical (user-based) evaluation. Expert evaluation is a relatively low-cost and efficient formative evaluation method applied even on system prototypes or design specifications up to the almost ready-to-ship product. The main idea is to present the tasks supported by the interface to an interdisciplinary group of experts who will take the part of would-be users and try to identify possible deficiencies in the interface design.

However, according to [9] "you can't really tell how good or bad your interface is going to be, without getting people to use it". This phrase expresses the broad belief that user-testing is inevitable in order to assess an interface. However, it is important to understand that test users can't tell us everything we might like to know, and that some of what they will tell us is useless. This is not done on purpose; for different reasons users often cannot give any reasonable explanation for what happened, or why they acted in a certain way. On the other hand, expert-based evaluations provide usually suboptimal results, as the evaluators try to "simulate" the user. So, a number of studies, such as [6]; [7]; [8], argues that the combination of an expert-based and a user-based approach provides always the best efficiency factor, by maximizing the evaluation outcomes while minimizing the needed resources. This study aims to propose a framework of such a combinatory evaluation in the domain of cultural heritage museum interfaces.

In this study, users and experts participated in the evaluation. The collected data was both of qualitative and quantitative form. This is not an unusual approach by so far; many studies report this structure. However, the preparation of the study and the elaboration techniques of the collected data are in many ways different from one study to another. In the present study, following approach has been set up, and is accordingly proposed.

1. A great deal of attention must be given so that the study conditions and influences are the same for both groups of the evaluators. In other words, the same evaluation approach must be used (observation, questionnaire, interview or other) for both groups, the same task and actions must be evaluated, and so on.
2. In the following stage a statistical elaboration of the quantitative data is performed. Its aim of this step is to define the "accordance factor". In other words, to determine whether the answers provided by the experts are in a statistical significant relation to the answers provided by the users. An independent samples t-test is the best approach in this case.
3. As next, the differences, and/or the factors where no statistical significant relation was encountered, must be pinpointed and recorded. Thus, two groups of results emerge: *coincident* opinions and *debatable* opinions.
4. The qualitative part of the evaluation comes now into consideration.
    a. For the coincident opinions, the personal meanings of both users and experts are interpreted, grouped and presented.
    b. For the debatable opinions, the differences are highlighted. Next, the personal meanings, which contribute to an understanding of the debate, must be presented in detail and the most serious differences must be emphasized.
5. A discussion must be made, according to the context of the study and a conclusion concerning the main question of the study must be stated.

## ARCO Evaluation

ARCO evaluation intended to be as participatory as possible, so it captured information by two homogeneous target groups of users [15]. One group comprised of the domain specialists, in our case the museum curators and the second group was the end-users representing the museum visitors that conducted a walkthrough of the system. As end-users, twenty nine participants were recruited from the University of Sussex, UK undergraduate and postgraduate population ranging from twenty to thirty years old. Ten domain experts took part in the evaluation aged between twenty-eight to sixty years old. All of them were museum curators from various departments of the Victoria and Albert Museum, London, UK. The end-users were not involved in the technical development of the ARCO system. The museum curators were involved in the technical development from an early stage setting user requirements and providing appropriate feedback during the early stages of implementation.

The main instrument used, as it is already mentioned, is the ACMA-ARIF Tutorial Questionnaire [2]. It provided a 1 to 5 Likert scale with space for additional comments. The participants have been informed about the ARCO system and the equipment that would be used, and were guided through a step-by-step process. The evaluation involved only one participant at a time and assistants instructed the museum curators and the end-users if they needed help. Neither track of the errors been done by the users, nor of the time needed to complete the tasks have not been kept, because it was not the research's intention to test the users' performance, but the system's performance.

The main evaluated exercise demonstrated how information about Cultural Objects can be presented in a form of an interactive scenario, where users can gain information not only by browsing it, but also by answering series of questions and resolving tasks. The interactive game corresponds to learning by doing and to what [3] calls *discovery learning*, where the users participate in an interactive experience and discover the correct answer. As a showcase the Fishbourne Roman Palace has been selected. To the welcome Web page appeared a brief story about the palace, a short introduction to the quiz that contained questions about the archaeological data and a summary of its scenario including its goals and rules. The users have set up the markers in the AR environment. A VRML model appeared on a marker and a question about an archaeological finding from the Fishbourne Roman Palace have to be answered correctly, by picking up and turning over the marker with the correct answer. Depending on whether the answer was correct or not, an appropriate virtual model appeared (a smiling or sad face) and a number of points for each correct answer could be scored. As a prize a virtual reconstruction of one of the palace wings was presented.

Six questions assessed this part of the evaluation:

1. Educational usefulness of the learning scenario within a museum/class room is (poor/excellent)
2. Presentation of questions in the AR environment is (poor/excellent)
3. Answering questions using double-sided markers is (very difficult/very easy)
4. Integration of the Web and AR presentation is (poor/excellent)
5. The scoring mechanism is (nonsense/essential)
6. Sounds accompanying the learning scenario are (nonsense/essential)

## Elaboration of the data collected

**Quantitative analysis.** According to the defined evaluation approach, the quantitative elaboration of the collected data has been performed.

Descriptive statistics showed that the Mean Values are relatively high, from 3,61 to 4,07. Maximum assessment was always at 5, while Minimum rarely at 1, usually at 2 and twice at 3. This indicates a good acceptance of the evaluated interface, at first view.

Main question for this part of the elaboration was whether the answers provided by the experts are in accordance to those provided by the users. So, an independent variable of nominal type has been set (expert or user). The dependent, measurable variables were accordingly the pending questions in the questionnaire. According to this classification, following hypotheses can be stated.

$H_0$: *There is no difference in the evaluators' opinions due to the fact that they belong to different groups.*

$H_a$: *There is difference in the evaluators' opinions due to the fact that they belong to different groups.*

In order to investigate these hypotheses, an independent samples t-test has been performed. The t-test showed that only in question 5 the homogeneity of the samples could be ensured, by means of an additional Levene's test. However, this was not important, as all significances of all questions are greater than the statistical significant value p=0,05, namely the are NOT statistical significant.

In order to strengthen this result, the non parametric tests Mann-Whitney and Wilcoxon have also been employed. The results were the same, the lowest value being at question 4 (p=0,091), yet also greater than the statistical acceptable limit of 0,05. So, these results can be considered as identical of those of the t-test, without threatening the validity of this claim.

So, as there was no statistical significant differentiation for all questions, the null hypothesis can not be rejected, so *there is no difference in the evaluators' opinions due to the fact that they belong to different groups*, and the whole sample (experts and users) can be considered as homogenous, as regards their estimations in this survey.

**Qualitative analysis.** The qualitative analysis consisted of the comments of the end-users and curators. The data collected by means of the ACMA-ARIF Tutorial Questionnaire have been grouped into three main categories of *positive comments*, *usability flaw characteristics* and *remarks/suggestions*. In the qualitative analysis the severity of the usability problems has also been taken into account. According to J. Nielsen, the usability problem is a combination of three factors, the *frequency* with which the problem occurs, which means if the problem is common or rare, the *impact* of the problem, which means if the problem is easy to be overcome or not and finally its *persistence*, if it is a problem that the users will repeatedly be bothered by it [11]. These factors have been thoroughly examined and provided useful information about the system.

## Results

A major area of usability problems is related with terminology and documentation problems. The qualitative analysis of the results revealed a severe usability problem mentioned several times by museum curators and end-users that the vocabulary used by the system creators was not always clear and familiar. For some users '*some of the specific terms were not self-explanatory and it needed some time to understand*'. Additionally, they needed clarification for terms, such as Web-Remote ARIF. Furthermore, there were difficulties with navigation around the system and its layout complexity. Four participants stated that the navigation to the system would be '*easier with instructions*', whereas five participants said that '*you need someone to guide you'*. These problems have been assessed as a severe obstacle, if someone wants to concentrate on an educational task.

Aesthetic issues about the *Learning Scenario* were the second issue of concern. There were comments about the fonts' size and colour, that were not very clear, and the evaluators suggested replacing the existing fonts with others that would have more contrast with their background.

The quality issues concerning the quality of the VRML models and the in most cases low-resolution images have been also discussed. Both textures of the 3D models and pictures of the cultural objects images had a low resolution, so as not to be 'heavy'. This choice was necessary for storage and transmission of the files mainly over the Internet. Both the museum curators and the end-users asked for more detailed images and higher resolution of the VRML models' textures.

A major finding of the usability test was that the users showed familiarity with characteristics used by Windows operational system and expected the same characteristics from ARCO system, so as to be easily recognised and understood. In the *Creation of Cultural Objects* and in the *Searching for Objects* domains, some comments revealed that difficulties encountered were related with the fact that some of the features of the ARCO system have differences from the Microsoft Windows. The users expected similar functions with the well-known operating system and consequently, they were disappointed when they have to interact with different system characteristics.

Some problems encountered with the functions of the system. In some cases, no standard functions were used. Three users found the use of shortcut Ctrl/Enter for saving the data was not intuitive. Another museum curator has noticed that '*it is not very common to use a right click in order to accomplish a task. It is usually done with menus'*. Certain users were confused by words that did not expect, such as 'finish' when the data are downloaded, instead of 'complete' that is used in Windows and in the *XDE Export and Import* task someone remarked that **'***The word 'Abort' should be replaced with the word 'Cancel'.* The comment that '*it is not very common to use a right click in order to accomplish a task'* also implies the Windows system, where all of the functionalities can be done either by menus, or by right click. In the section where there were questions that have to do with the *Creation of a Working Space* of a virtual museum exhibition, they proposed that a button with an ok or a save will be better than simply close the window, because it is not clear if the changes have been saved. In the *Refinement of Cultural Objects* '*The word 'Finish' used when the data has finished loading could be replaced by the word 'Complete'*.

Most of the proposed ideas about the improvement of the system were related with the Learning Scenario. The interviewers have asked to add a brief introduction to the quiz '*for example a clip or a movie'* and provide more explanation about the scoring mechanism, and proposed the use of a clapping hand instead of a smiley face to indicate a correct answer. Voice recognition of answers In addition to this, museum curators have asked to omit the sounds of applause and the smiling faces when the user selects a correct answer, as well as the sounds of disappointment and the sad faces used when the wrong choice is selected.

## Conclusions

In conducting this study, many interesting themes emerged from the data collected that were related with terminology and documentation design, quality consistency and standards of GUI interfaces issues. The evaluation triggered a series of system refinements. According to the museum curators' and end-users' feedback changes have been made to the technical problems that have been encountered during the experimental procedure. Bugs of the system have been fixed and problematic features related with the interactive quiz have been arranged accordingly.

As regards the defined evaluation framework, the general impression is that the flawless performance of the qualitative part of the study and the easy elaboration of the findings (grouping, assessing the severity, refining), support the claim that the predefined combinatory evaluation framework, as defined in section 3, is applicable and useful. Aim of its design was to constitute a fair and robust environment, in order not to threaten the validity of the results, and in particular the comparison of the users' opinions to those of the experts', an issue that provokes many debates in literature. Although this framework is rather simple and intuitive, it is rarely adhered to, as in most reported studies of this kind something is missing. In most cases, either the integrated design of the study or the statistical elaboration is neglected. However, sidesteps from this procedure may threaten the validity of the results, as experts and users

have usual very different views and very different expertise as regards to the same entity under evaluation.

Further research on the domain could take into its scope the generalization of this result and the validation of this framework in other domains as well. The practical importance of such an approach is more than obvious, as a well defined evaluation framework utilizing users as well as experts and elaborating quantitative as well as qualitative data, would be applicable in a substantial broader context of educational interfaces, providing thus a fair methodology in order to elicit reliable and valid evaluation results.

## Acknowledgments

Part of this work has been funded by the Marie Curie Actions Human resources and Mobility Marie Curie training site: Virtual Reality and computer graphics, project HPMT-CT-2001-00326. Also, the authors would like to thank Asma Almosawi for her contribution to the interviewing procedure and the preliminary results of the research and Katerina Mania for her suggestions during the first phase of the research.

## References

[1] ARCO, Augmented Representation of Cultural Objects (2004). http://www.arco-web.org/ (Retrived 7 Oct 2004).

[2] ARCO (2005). Evaluation Report, D16: Assessment and Evaluation report on the ARCO system and its components, http://www.arco-web.org/TextVersion/Documents/Deliverables/ARCO-D16-R-1.0-170904.pdf (Retrieved 20 Dec. 2005).

[3] Hein, George E., (1998). Learning in the museum. London: Routledge.

[4] Jones J. and Christal M., (2002). *The Future of Virtual Museums: On-Line, Immersive, 3D Environments*. Created Realities Group.

[5] ISO 9241 - International Standarization Organization. (1998). *Ergonomic requirements for office work with visual display terminals (VDT's)*.

[6] Kantner L., Rosenbaum S., (1997). Usability Studies of WWW sites: Heuristic Evaluation vs Laboratory Testing.

[7] Karoulis, A., and Pombortsis, A. (2000). Evaluating the Usability of Multimedia Educational Software for Use in the Classroom Using a «Combinatory Evaluation» Approach. Proc. of EDEN 4th Open Classroom Conference, 20-21 Nov 2000, Barcelona, Spain

[8] Karoulis, A., Demetriades, S., and Pombortsis, A. (2005). Comparison of Expert-Based and Empirical Evaluation Methodologies in the Case of a CBL Environment: The «Orestis» Experience. *Computers and Education*. (to appear).

[9] Lewis, C. and Rieman, J. (1994). *Task-centered User Interface Design - A practical introduction*, 1994. ftp.cs.colorado.edu/pub/cs/distribs/HCI-Design-Book (Retrived 20 May 2001).

[10] Nielsen J., (1993). *Usability Engineering,* New York, NY: Academic Press Inc.

[11] Nielsen J., (2005). *Severity Ratings for Usability Problems*, http://www.useit.com/papers/heuristic/severityrating.html (retrieved 20 Dec 2005).

[12] Patel M., Walczak K., White M. and Cellary W., (2003). Digitisation to Presentation-Building Virtual Museum Exhibitions. *Proceedings of the International Conference on Vision, Video and Graphics*, Bath, UK, pp. 189-196.

[13] Preece, J. Roger, Y. Sharp, H. Beyon, D. Holland, S. & Carey, I. (1994). *Human Computer Interaction.* Wokingham, UK: Addison Wesley Publishing.

[14] Scali G., Segbert M., Morganti B., 2002, *Multimedia applications for innovation in cultural heritage: 25 European trial projects and their accompanying measure TRIS*. In 68th IFLA Council and General Conference August 18-24, 2002.

[15] Sylaiou S., Almosawi A., Mania K., White M., (2004). Preliminary Evaluation of the Augmented Representation of Cultural Objects System, in *Proceedings of the 10th International Conference on Virtual Systems and Multimedia, Hybrid Realities-Digital Partners, Explorations in Art, Heritage, Science and the Human Factor, VSMM 2004 Conference*, 17-19 November 2004, Softopia Japan, Ogaki City, Japan, pp. 426-431, ISBN 4-274-90634-5 (Ohmsha, Ltd.), ISBN 1-58603-481-2 (IOS Press).

[16] Wojciechowski R., Walczak K., White M. and Cellary W., (2004). Building Virtual and Augmented Reality Museum Exhibitions. *Proceedings of the Web3D 2004 Symposium – the 9th International Conference on 3D Web Technology*, *ACM SIGGRAPH*, Monterey, California (USA), pp. 135-144.